**短语相似度评定数据集说明**

1 如何确定短语种类

由于对单词含义进行整合很大程度上是为了理解句子的含义，一般认为句子含义是由句法树中节点自下而上进行组合，所以我们选择句法树中由叶子节点构成的高频子树为候选的短语种类。由于中文的单字词歧义数量多，因此我们只考虑双字词。

通过对chtb标准树库进行统计，我们可以得到前三个高频短语分别为：‘名词+名词’、‘基数+量词’和‘形容词+名词’。由于‘基数+量词’多为抽象词，且短语之间很难评定相似度，因此我们不考虑这种短语。英文短语相似度数据集中短语类型除了‘名词+名词’和‘形容词+名词’之外，还有‘动词+宾语’类型。考虑到‘动词+宾语’类型在句子中较为常见，只是短语中间多含形容词等修饰成分，因此在chtb树库中没有被统计，因此我们将‘动词+宾语’类型加入评定数据种类。中文‘的’在文本中使用频率非常高，而且‘动词+的+名词’类型的短语使‘动词+名词’的短语含义有所改变，使‘动词+名词’由动词短语变为名词短语，因此我们将这种类型也纳入评定短语种类。

2 如何抽取短语

我们选择搜狗新闻语料和百度百科语料，共13G大小，作为模型的训练集。我们在训练集中抽取短语相似度的评定语料，这样可以用评定语料对模型训练结果进行更好的评价。

首先对纯文本语料进行分词和词性标注，对于‘名词+名词’和‘形容词+名词’短语，我们直接在词性标注的语料中进行抽取；对于‘动词+宾语’，针对汉语文本使用情况，我们设计规则对其进行抽取。抽取规则为：在动词到标点符号间名词都做为动词的候选宾语，最后选择高频短语和人工筛选会去除非‘动词+宾语’的假短语。最终得到候选‘名词+名词’短语12958552个；‘动词+宾语’短语61038730个；‘形容词+名词’短语2379905个；‘动词+的+名词’短语3823599个。

3 如何确定评定短语对

得到三种类型的短语候选集合后，我们需要确定短语对，让被试对短语对进行相似度评定。

首先，两个短语交换单词后得到两个新的的短语，选择这两个新的短语出现在短语候选集合且为高频的短语，比如‘文学 作品’和‘文艺 创作’，交换单词后得到‘文学 创作’和‘文艺 作品’。为了缩小短语对候选集合的规模，我们认为这样选出的短语对在一定程度上是相似的。这样得到候选‘名词+名词’短语79702对；‘动词+宾语’短语554456对；‘形容词+名词’短语27904对；‘动词+的+名词’短语22984对。

接着，对于每个短语，只保留与其最相似的一个短语形成的短语对，并且对于交换顺序得到的短语对，只保留如‘文学作品’和‘文学创作’中的高频短语对。利用同义词词林对剩下的短语对进行相似度评定，选择相似度最高的前200个短语对，并且为了排除自动标注工具的错误，对人工标注的人民日报数据集抽取4种类型的短语，选择在人民日报短语集合中出现的短语作为下一步候选短语对。这样得到候选‘名词+名词’短语56对；‘动词+宾语’短语86对；‘形容词+名词’短语22对；‘动词+的+名词’短语46对。由于‘形容词+名词’和‘动词+的+名词’短语对的候选数量较少，因此我们分别选取没有在人民日报中出现的30个高频‘形容词+名词’短语对和10个高频‘动词+的+名词’短语对加入到候选集合中。

最后，人工删除假短语（对于‘名词+名词’短语，假短语为没有在搜索引擎中形成短语链接的短语；对于其他短语类型，根据连贯程度等指标进行判断），如‘国际质量’、‘拥有公司’等。最终得到‘名词+名词’短语41对；‘动词+宾语’短语35对；‘形容词+名词’短语41对；‘动词+的+名词’短语40对。’动词+宾语’短语数量较少，因此从高频短语对中选择最高频的且非假短语的5对加入数据集。保留每种类型短语对的最高频40对短语对，令每种类型短语评价集数量一致。

4 如何确定最终评定短语对组合

上述短语对都是同义词林统计出的高相似度短语对，为了让评定数据集可以反映高相似度、中等相似度和低相似度三种类型，我们对上述40对短语对进行重组，对重组后短语对进行同义词词林相似度计算，选择相似度最高的40对作为中等相似度的短语对，选择相似度最低的40对作为的相似度的短语对。这样我们就得到了每种数据类型120对短语的相似度评定数据集。

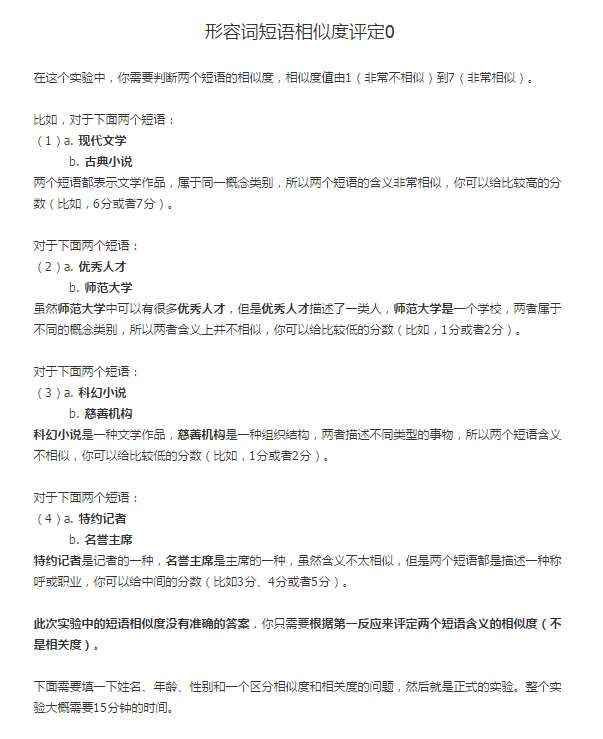
5 如何进行人工评定及评价标准

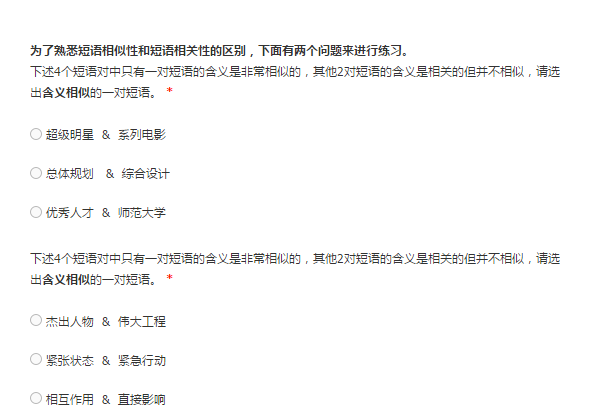
由于词义相似度和词义相关度是两个不同的指标，如‘艺术院校’和‘工程院校’相似，因为同属于学校的范围；‘艺术院校’和‘音乐创作’相关但不相似，因为不属于同一种概念类别，但是会经常出现在同一场景中。我们想要对短语的相似性进行评定，就要在实验的指导语中，对这两种指标进行说明，并且引导被试对短语的相似性进行评定。我们选择分组的方式对短语进行评定，即每个页面中出现一组9个短语对，下一个页面的第一个短语对为上一个页面出现过的短语对，这样被试可以在评定过程中有参照短语对。实验中在每种类型短语评定开始时，为了让被试熟悉相似性和相关性的差别，还增加了一页‘检查点’，页面上为一组抽象短语的和一组具体短语，被试需要在3个选项中选择和这个短语最相似的一个短语，选项中只有一个是非常相似的短语，其他2个为相关短语。

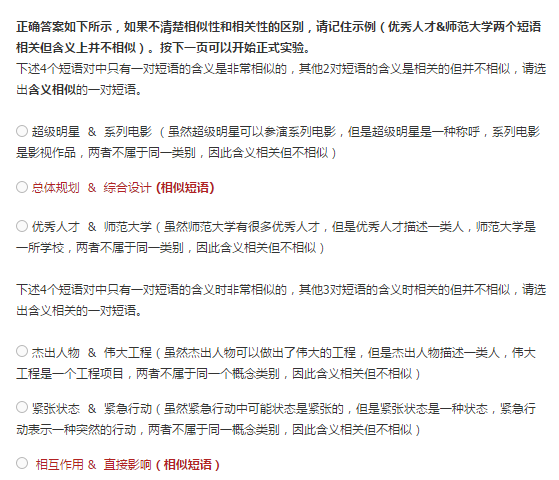
**数据采集：**

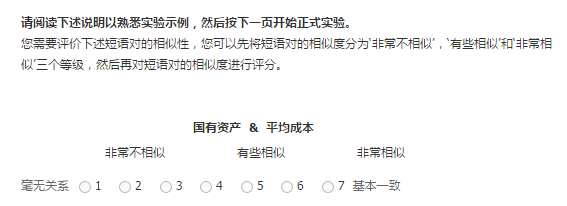
实验问卷由腾讯问卷系统（<http://wj.qq.com/index.html>）采集，通过QQ进行传播，共有约150人（其中有些人参与了多份问卷）有偿参与实验，被试的母语均为汉语，且均不是语言学相关专业。实验共收集有效问卷178份，其中，名词短语评定数据40份，动词短语评定数据54份，形容词短语评定数据42份，动词+的+名词短语42份。经过人工检查，筛除相关度较低的数据，剩余名词短语评定数据39份，动词短语评定数据46份，形容词短语评定数据40份，动词+的+名词短语41份。实验被试的男女比例为3:1，年龄为集中在20~30岁左右。

**实验指导语与说明：**









**数据分析：**

首先，我们通过前10个最相似短语，直观的看评定数据的准确性：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Adjective-noun | Noun-noun | | Verb-object | | Verb-de-noun |
| |  | | --- | | 有关\_资料&相关\_材料 | | 有关\_法律&相关\_法规 | | 有关\_人士&相关\_人员 | | 有关\_规定&相关\_条款 | | 有关\_事项&相关\_事宜 | | 有关\_部门&相关\_单位 | | 相关\_企业&有关\_公司 | | 突出\_贡献&杰出\_成就 | | 重要\_角色&主要\_人物 | | 合法\_权益&正当\_利益 | | |  | | --- | | 国家\_主席&政府\_首脑 | | 专业\_技术&职业\_技能 | | 国际\_金融&全球\_经济 | | 书画\_艺术&绘画\_作品 | | 工业\_园区&产业\_基地 | | 市场\_价格&商品\_价格 | | 经济\_体制&经济\_结构 | | 经济\_体制&金融\_体系 | | 地理\_环境&地质\_条件 | | 市委\_常委&省委\_党校 | | |  | | --- | | 出席\_会议&参加\_大会 | | 召开\_会议&举行\_大会 | | 取得\_成绩&获得\_成果 | | 成立\_公司&建立\_企业 | | 分析\_习题&解析\_试题 | | 促进\_科技&推动\_科学 | | 了解\_情况&掌握\_信息 | | 采用\_方式&使用\_方法 | | 促进\_发展&推动\_进步 | | 得到\_支持&获得\_认可 | | | |  | | --- | | 喜欢\_的\_歌手&喜爱\_的\_歌星 | | 发布\_的\_信息&公布\_的\_消息 | | 喜欢\_的\_食物&喜爱\_的\_食品 | | 采用\_的\_方法&采取\_的\_方式 | | 取得\_的\_收入&获得\_的\_收益 | | 得到\_的\_消息&获得\_的\_信息 | | 上升\_的\_趋势&上涨\_的\_势头 | | 造成\_的\_后果&导致\_的\_结果 | | 喜欢\_的\_音乐&喜爱\_的\_歌曲 | | 公布\_的\_消息&发布\_的\_公告 | | |

接着，利用一系列Kruskal-Wallis rank sum 试验来检测我们之前使用的三个band中的数据和人工评定的相似度值。在每种类型的短语对中，被试在三个band中的打分是显著不同的（p<.01）。下表展示了不同数据类型短语对在三个band中的均值，标准差和标准误。由均值（High>Meduim>Low）可以看出，我们之前利用同义词词林划分的三个band是有效的。

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Adjective-noun | | | Noun-noun | | | Verb-object | | | Verb-de-noun | | |
|  | M | SD | SE | M | SD | SE | M | SD | SE | M | SD | SE |
| High | 4.46 | 1.859 | 0.046 | 4.02 | 1.738 | 0.044 | 3.94 | 1.902 | 0.044 | 4.72 | 1.981 | 0.049 |
| Medium | 2.77 | 1.672 | 0.042 | 2.79 | 1.758 | 0.045 | 2.60 | 1.570 | 0.037 | 2.65 | 1.759 | 0.043 |
| Low | 1.48 | 0.830 | 0.021 | 1.87 | 1.241 | 0.031 | 1.53 | 0.895 | 0.021 | 1.86 | 1.232 | 0.030 |

我们还检验了被试在不同短语类型中表现的一致程度。人工评定数据中被试间的一致度是我们模型的一个上限，可以让模型得出的结果与人作比较。为了计算被试间的一致性，我们使用留一交叉验证。对于每种短语类型，我们将被试分为两组，一组为一个被试，另一组为其他被试（对多个被试求平均），我们计算这两组数据间的spearson相关性系数ρ，重复多次，直到每个人都被单独分为一组。得到结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Spearman correlation | Noun-Noun | Verb-Noun | Adjective-Noun | Verb-de-noun |
| Mean | 0.7836 | 0.817707 | 0.836137 | 0.839702 |
| Max | 0.904774 | 0.915898 | 0.91593 | 0.932505 |
| Min | 0.589749 | 0.608979 | 0.562192 | 0.567367 |
| SD | 0.073536 | 0.066712 | 0.065086 | 0.08256 |

结果表示，虽然被试认为短语相似度的评定任务很难，但是被试间的结果有很高的一致性。名词-名词短语的相似性评定比较困难，形容词-名词和动词-的-名词的短语相似性评定比较简单。原因可能是，名词-名词组合成的短语包含了两个单词共同的含义，相比形容词-名词或者动词-的-名词短语中前面的词为后面名词的修饰成分，主要由后面的名词形成短语的主要概念，需要更多的认知负荷，因而更难。